BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e36095
doi: 10.3897/biss.3.36095

OPEN ACCESS

Conference Abstract

SPECIES: Supporting big-data-driven research

Raul Sierra-Alcocer[‡,§], Christopher R Stephens[§,|], Juan M Barrios[‡], Constantino González-Salazar[¶,§], Juan Carlos Salazar Carrillo[‡], Pedro Romero Martínez[‡]

‡ Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO), Mexico City, Mexico
§ Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Mexico City, Mexico
| Instituto de Ciencias Nucleares, Universidad Nacional Autónoma de México, Mexico City, Mexico
¶ Departamento de Ciencias Ambientales, Universidad Autónoma Metropolitana, Lerma de Villada, Mexico

Corresponding author: Raul Sierra-Alcocer (raul.sierra@conabio.gob.mx)

Received: 10 May 2019 | Published: 21 Jun 2019

Citation: Sierra-Alcocer R, Stephens C, Barrios J, González-Salazar C, Salazar Carrillo J, Romero Martínez P (2019) SPECIES: Supporting big-data-driven research. Biodiversity Information Science and Standards 3: e36095. https://doi.org/10.3897/biss.3.36095

Abstract

SPECIES (Stephens et al. 2019) is a tool to explore spatial correlations in biodiversity occurrence databases. The main idea behind the SPECIES project is that the geographical correlations between the distributions of taxa records have useful information. The problem, however, is that if we have thousands of species (Mexico's National System of Biodiversity Information has records of around 70,000 species) then we have millions of potential associations, and exploring them is far from easy. Our goal with SPECIES is to facilitate the discovery and application of meaningful relations hiding in our data.

The main variables in SPECIES are the geographical distributions of species occurrence records. Other types of variables, like the climatic variables from WorldClim (Hijmans et al. 2005), are explanatory data that serve for modeling. The system offers two modes of analysis. In one, the user defines a target species, and a selection of species and abiotic variables; then the system computes the spatial correlations between the target species and each of the other species and abiotic variables. The request from the user can be as small as comparing one species to another, or as large as comparing one species to all the species in the database. A user may wonder, for example, which species are usual neighbors of the jaguar, this mode could help answer this question. The second mode of analysis gives a network perspective, in it, the user defines two groups of taxa (and/or environmental variables), the output in this case is a correlation network where the weight

© Sierra-Alcocer R et al. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

of a link between two nodes represents the spatial correlation between the variables that the nodes represent. For example, one group of taxa could be hummingbirds (Trochilidae family) and the second flowers of the Lamiaceae family. This output would help the user analyze which pairs of hummingbird and flower are highly correlated in the database. SPECIES data architecture is optimized to support fast hypotheses prototyping and testing with the analysis of thousands of biotic and abiotic variables. It has a visualization web interface that presents descriptive results to the user at different levels of detail.

The methodology in SPECIES is relatively simple, it partitions the geographical space with a regular grid and treats a species occurrence distribution as a present/not present boolean variable over the cells. Given two species (or one species and one abiotic variable) it measures if the number of co-occurrences between the two is more (or less) than expected. If it is more than expected indicates a signal of a positive relation, whereas if it is less it would be evidence of disjoint distributions.

SPECIES provides an open web application programming interface (API) to request the computation of correlations and statistical dependencies between variables in the database. Users can create applications that consume this 'statistical web service' or use it directly to further analyze the results in frameworks like R or Python. The project includes an interactive web application that does exactly that: requests analysis from the web service and lets the user experiment and visually explore the results. We believe this approach can be used on one side to augment the services provided from data repositories; and on the other side, facilitate the creation of specialized applications that are clients of these services. This scheme supports big-data-driven research for a wide range of backgrounds because end users do not need to have the technical know-how nor the infrastructure to handle large databases.

Currently, SPECIES hosts: all records from Mexico's National Biodiversity Information System (CONABIO 2018) and a subset of Global Biodiversity Information Facility data that covers the contiguous USA (GBIF.org 2018b) and Colombia (GBIF.org 2018a). It also includes discretizations of environmental variables from WorldClim, from the Environmental Rasters for Ecological Modeling project (Title and Bemmels 2018), from CliMond (Kriticos et al. 2012), and topographic variables (USGS EROS Center 1997b, USGS EROS Center 1997a). The long term plan, however, is to incrementally include more data, specially all data from the Global Biodiversity Information Facility. The code of the project is open source, and the repositories are available online (Front-end, Web Services Application Programming Interface, Database Building scripts).

This presentation is a demonstration of SPECIES' functionality and its overall design.

Keywords

big data, spatial data mining, correlation analysis, species occurrence records, species distributions

Presenting author

Raúl Sierra-Alcocer

Presented at

Biodiversity_Next 2019

Conflicts of interest

None

References

- CONABIO (2018) Sistema Nacional de Información sobre Biodiversidad. Registros de ejemplares. 25/07/2018. Comisión Nacional para el Conocimiento y Uso de la Biodiversidad. Release date: 2018-7-25. URL: https://doi.org/10.15468/dl.ikfkok
- GBIF.org (2018a) GBIF Occurrence Download. Release date: 2018-6-12. URL: https://doi.org/10.15468/dl.menapb
- GBIF.org (2018b) GBIF Occurrence Download. Release date: 2018-5-07. URL: https://doi.org/10.15468/dl.iixcqe
- Hijmans R, Cameron S, Parra J, Jones P, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25 (15): 1965-1978. https://doi.org/10.1002/joc.1276
- Kriticos D, Webber B, Leriche A, Ota N, Macadam I, Bathols J, Scott J (2012) CliMond: global high-resolution historical and future scenario climate surfaces for bioclimatic modelling. Methods in Ecology and Evolution 3 (1): 53-64. https://doi.org/10.1111/j.2041-210x.2011.00134.x
- Stephens CR, Sierra-Alcocer R, González-Salazar C, Barrios JM, Salazar Carrillo JC, Robredo Ezquivelzeta E, Del Callejo Canal E (2019) SPECIES: A platform for the exploration of ecological data. Ecology and Evolution 9 (4): 1638-1653. https://doi.org/10.1002/ece3.4800
- Title P, Bemmels J (2018) ENVIREM: an expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling. Ecography 41 (2): 291-307. https://doi.org/10.1111/ecog.02880
- USGS EROS Center (1997a) USGS 30 ARC-second Global Elevation Data, GTOPO30. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. URL: https://rda.ucar.edu/datasets/ds758.0/
- USGS EROS Center (1997b) HYDRO1k Elevation Derivative Database. URL: https://doi.org/10.5066/F77P8WN0